

ESTATÍSTICA NÃO-PARAMÉTRICA

Aula 5

Prof. Dr. Edmilson Rodrigues Pinto
Faculdade de Matemática - UFU
edmilson@famat.ufu.br

1

Caso de k amostras relacionadas

O objetivo principal desses testes é comprovar a hipótese de que k amostras tenham sido extraídas da mesma população ou de populações idênticas.

2

Nesse caso, serão abordadas técnicas não-paramétricas de comparação de 3 ou mais grupos relacionados entre si, onde os dados têm a seguinte estrutura

Blocos	Grupos (Tratamentos)			
	G_1	G_2	\dots	G_k
B_1	x_{11}	x_{12}	\dots	x_{1k}
\vdots	\vdots	\vdots	\vdots	\vdots
B_n	x_{n1}	x_{n2}	\dots	x_{nk}

Onde os blocos representam as unidades amostrais utilizadas no experimento e os tratamentos são as k condições de avaliação.

3

Teste Q de Cochran

É uma extensão do teste de McNemar para duas amostras e se aplica na verificação de diferença entre três ou mais grupos de frequências ou proporções.

Os dados devem estar em escala nominal ou ordinal dicotômica.

O teste envolve $k \geq 2$ tratamentos que são aplicados independentemente para cada um de n indivíduos. Os resultados de cada tratamento são guardados, como uma variável dicotômica de sucesso ou fracasso (uns e zeros), em uma tabela de contingência.

O teste de Cochran permite investigar se existe diferença entre os k tratamentos

4

Hipóteses

H_0 : não existe diferença entre os efeitos dos k tratamentos.

H_1 : existe diferença entre, pelo menos, dois tipos de tratamentos.

Procedimento do teste

1. Atribua o valor 1 (um) para cada sucesso e o valor 0 (zero) para cada fracasso.
2. Disponha os dados em uma tabela de contingência com n linhas e k colunas. (n : nº de casos e k : nº de tratamentos).

5

3. Determine o valor de Q , dado por:

$$Q = \frac{(k-1) \left[k \sum_{j=1}^k G_j^2 - \left(\sum_{j=1}^k G_j \right)^2 \right]}{k \sum_{i=1}^n L_i - \sum_{i=1}^n L_i^2}$$

Onde: G_j é a soma dos valores (1 ou 0) das j colunas. $j = 1, \dots, k$

L_i é a soma dos valores (1 ou 0) das i linhas. $i = 1, \dots, n$

$$Q \sim \chi_{k-1}^2$$

6

Decisão

Para um nível de significância α , obtenha $\chi_{Tab}^2 = \chi_{k-1, \alpha}^2$

Se $Q \geq \chi_{Tab}^2$, rejeite H_0 , caso contrário, aceite.

Ou calcule o p-valor ($\hat{\alpha}$), onde $\hat{\alpha} = P(\chi_{k-1}^2 \geq Q)$

Se $\hat{\alpha} < \alpha$ rejeite H_0 , caso contrário, aceite.

7

Exemplo Cada um dos quatro fãs de futebol criou um sistema para prever os resultados dos jogos da 1ª divisão do campeonato brasileiro. Foram escolhidos, ao acaso, seis jogos e cada um dos fãs previu o resultado de cada jogo. Os resultados dos prognósticos foram dispostos em uma tabela de contingência, utilizando 1 (um) para prognóstico bem sucedido e 0 (zero) para prognóstico falhado. Queremos testar a hipótese de que cada um dos fãs tem um sistema de igual efeito para prever o resultado dos jogos a um nível de 5% de significância. Os dados são mostrados na tabela seguinte.

8

Jogos	Fãs				Total L_i
	1	2	3	4	
1	1	1	0	0	2
2	1	0	1	0	2
3	1	1	1	0	3
4	0	1	1	1	3
5	0	1	0	0	1
6	1	1	0	1	3
Total G_j	4	5	3	2	14

9

Teste de Friedman

É útil quando deseja-se comprovar a hipótese de que as k amostras relacionadas provêm da mesma população.

Neste tipo de estudo observa-se o mesmo grupo de indivíduos sob cada uma das k condições, ou então formam-se conjuntos de indivíduos homogêneos entre si e estes são alocados aleatoriamente a cada uma das condições.

O teste pode ser considerado como uma extensão do teste do sinal para comparação de duas amostras pareadas.

Exige-se que os dados estejam, pelo menos, em escala ordinal.

10

Hipóteses

H_0 : as k amostras relacionadas provêm da mesma população (ou de populações com a mesma mediana)

H_1 : as k amostras relacionadas provêm de populações distintas.

ou

H_0 : $\mu_1 = \dots = \mu_k$

H_1 : $\mu_i \neq \mu_j$ para, pelo menos, um $i \neq j$ onde $i, j = 1, \dots, k$

11

Procedimento do teste

1. Disponha os dados em uma tabela de contingência com n linhas e k colunas.
2. Atribua postos de 1 a k aos valores de cada linha.
3. Determine a soma dos postos de cada coluna, R_j , com $j = 1, \dots, k$
4. Calcule o valor da estatística

$$F_r = \frac{12}{nk(k+1)} \sum_{j=1}^k R_j^2 - 3n(k+1)$$

Onde n : é o número de linhas

k : é o número de colunas

R_j : é a soma dos postos na coluna j .

12

Decisão

O método para determinar a probabilidade de ocorrência, sob a hipótese nula, associado a um valor igual ou mais extremo do que o valor observado F_r , depende dos tamanhos de n e k .

- **Caso 1:** para $k = 3$ com n de 2 a 9 e para $k = 4$ com n de 2 a 4, use a Tabela N (apostila) para obter as probabilidades exatas associadas a valores tão grandes quanto F_r .

- **Caso 2:** caso k ou n excedam os valores da Tabela N, podemos usar a aproximação para a F_r , ou seja, $F_r \sim \chi_{k-1}^2$

Se o p-valor, obtido pelo método adequado (caso 1 ou 2) não superar α rejeite H_0 , caso contrário, aceite.

13

Exemplo

A fim de avaliar se houve mudança no aprendizado de seus alunos, um professor reteve as médias de um grupo de 4 alunos no final de cada trimestre, obtendo a seguinte tabela.

Alunos	Trimestre		
	1º	2º	3º
A	8 (1)	14 (2)	15 (3)
B	15 (1)	17 (2,5)	17 (2,5)
C	11 (1)	13 (2)	14 (3)
D	7 (1)	10 (2)	12 (3)
\bar{R}_j			

14

Comentário: (Teste das comparações múltiplas)

O teste de Friedman diz apenas (caso H_0 seja rejeitada) que os tratamentos são diferentes. Se quiséssemos verificar se existe diferença entre dois tratamentos quaisquer, o procedimento usado é o teste de comparações múltiplas, que é um complemento do teste de Friedman.

$$H_0: \mu_i = \mu_j$$

$$H_1: \mu_i \neq \mu_j \quad \text{para } j = 1, \dots, i-1, i+1, \dots, k$$

15

Procedimento

1. Calcula-se, para cada par de tratamentos, $|R_i - R_j|$ onde i é fixo (tratamento que ser comparar) e $j = 1, \dots, i-1, i+1, \dots, k$

R_j é a soma dos postos do tratamento j

R_i é a soma dos postos do tratamento i

2. Obtenha $d = \min \{|R_i - R_j|; j = 1, \dots, i-1, i+1, \dots, k\}$

3. Calcule

$$Z_{\alpha/k(k-1)} \sqrt{\frac{nk(k+1)}{6}}$$

Onde $Z_{\alpha/k(k-1)}$ representa o quantil $\alpha/k(k-1)$ da distribuição normal padrão.

16

Decisão

Se $d \geq Z_{\alpha/k(k-1)} \sqrt{\frac{nk(k+1)}{6}}$ rejeite H_0 , caso contrário, aceite.

17

Caso de k amostras independentes

O objetivo é verificar se diversas variáveis independentes devem ser consideradas como provenientes da mesma população.

18

Teste de Kruskal-Wallis

O objetivo é verificar se as diferentes k amostras independentes provêm da mesma população ou de populações com a mesma mediana.

O teste supõe que a variável tenha distribuição contínua e exige mensuração no mínimo ao nível ordinal.

Este teste é uma extensão do teste de Wilcoxon para duas amostras independentes e se utiliza de postos atribuídos aos valores observados.

19

Hipóteses

H_0 : as k amostras independentes provêm da mesma população (ou de populações com a mesma mediana)

H_1 : as k amostras independentes provêm de populações distintas.

ou

H_0 : $\mu_1 = \dots = \mu_k$

H_1 : $\mu_i \neq \mu_j$ para, pelo menos, um $i \neq j$ onde $i, j = 1, \dots, k$

20

Procedimento do teste

1- Disponha, em postos, as observações de todos os k grupos (amostras) numa única série, atribuindo-lhes postos de 1 a n . Onde $n = n_1 + \dots + n_k$ e n_j é o tamanho da amostra j (do grupo j).

2- Determine o valor de R_j , a soma dos postos da amostra j , para $j = 1, \dots, k$

3- Caso não haja empates nos postos, calcule o valor de H como:

$$H = \frac{12}{n(n+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} - 3(n+1)$$

Onde k é o número de amostras, $n = \sum_{j=1}^k n_j$ é o número total de casos, em todas as amostras e n_j é o tamanho da amostra j .

21

4- caso haja empates nos postos, atribua a cada um dos postos empatados a média desses postos caso não houvesse empates. Neste caso, o valor de H é influenciado pelos empates e precisa ser corrigido. Assim,

$$H = \frac{\frac{12}{n(n+1)} \sum_{j=1}^k R_j^2 - 3(n+1)}{1 - \sum_{i=1}^m T_i / (n^3 - n)}$$

Onde $T_i = t_i^3 - t_i$, m é o número de grupos de observações empatadas e t_i é o número de observações empatadas em cada grupo.

22

Decisão

O método para determinar a probabilidade de ocorrência, sob a hipótese nula, associado a um valor igual ou mais extremo do que o valor observado H depende dos tamanhos de n e k .

- **Caso 1:** para $k = 3$ e $n_1, n_2, n_3 \leq 5$, use a Tabela **O** (apostila) para obter, sob H_0 , a probabilidade exata associada a um H tão grande quanto o observado.

- **Caso 2:** para outros valores de k , a distribuição de H pode ser aproximada por uma distribuição qui-quadrado, ou seja, $H \sim \chi_{k-1}^2$

Se o p-valor, obtido pelo método adequado (caso 1 ou 2) não superar α rejeite H_0 , caso contrário, aceite.

23

Exemplo

Numa pesquisa sobre qualidade de vinho, três tipos de vinho foram avaliados por quatro degustadores. Cada degustador provou 12 amostras (4 de cada tipo) e atribuiu a cada uma delas uma nota de zero a dez. As médias das notas atribuídas pelos 4 degustadores, a cada uma das amostras, foram as seguintes.

Degustadores	Vinho		
	Tipo 1	Tipo 2	Tipo 3
A	5,0 (1)	8,3 (7)	9,2 (11)
B	6,7 (2)	9,3 (12)	8,7 (9)
C	7,0 (4)	8,6 (8)	7,3 (5)
D	6,8 (3)	9,0 (10)	8,2 (6)

Com base nas médias das notas fornecidas pelos degustadores, verifique se há diferença entre os tipos de vinho.

24

Medidas de correlação

Coefficiente de correlação por postos de Kendall

Denominamos $-1 \leq \tau \leq 1$ o coeficiente de correlação por postos de Kendall. O objetivo é medir a intensidade da correlação dos postos entre duas variáveis (grupos) X e Y .

Se a correspondência entre os postos for perfeita, no sentido positivo, ou seja, se todos os postos forem iguais para as duas variáveis, $\tau = +1$ (correlação perfeita positiva). Se houver uma discordância perfeita entre os postos das duas variáveis, ou seja, se o primeiro posto de uma corresponder ao último posto da outra, se o segundo posto de uma corresponder ao penúltimo da outra e assim por diante, $\tau = -1$ (correlação perfeita negativa)

25

Procedimento

Sejam os pares $(x_1, y_1), \dots, (x_n, y_n)$

1- Atribua postos de 1 a n para cada uma da variáveis X e Y , de modo a formar n pares de postos.

2- Ordene os pares de postos de acordo com os postos de X . Desta forma, os postos de Y , ordenados de acordo com os postos de X , serão: $PY_{x_1}, PY_{x_2}, \dots, PY_{x_n}$

3- Usando $PY_{x_1}, PY_{x_2}, \dots, PY_{x_n}$ calcule $P = \sum_{i=1}^n p_i$

onde p_i representa o número de postos à direita de PY_{x_i} maiores do que ele.

26

4- Calcule

$$S = 2P - \frac{n(n-1)}{2}$$

5- O coeficiente de correlação por postos de Kendall é obtido então por:

$$\tau = \frac{2S}{n(n-1)}$$

27

Exemplo

Suponhamos que um número de alunos está classificado por postos de acordo com suas habilidades em matemática e em música. A tabela seguinte mostra os valores de cada aluno.

	Aluno									
Disciplina	A	B	C	D	E	F	G	H	I	J
Matemática (X)	7	4	3	10	6	2	9	8	1	5
Música (Y)	5	7	3	10	1	9	6	2	8	4

	Aluno									
Disciplina	I	F	C	B	J	E	A	H	G	D
Matemática (X)	1	2	3	4	5	6	7	8	9	10
Música (Y)	8	9	3	7	4	1	5	2	6	10

28

Comentários

1- Em caso de haver observações empatadas, atribuímos a elas a média dos postos que lhe caberiam se não houvesse empate. Nesse caso, o valor de τ seria modificado da seguinte forma

$$\tau = \frac{S}{\sqrt{\frac{1}{2}n(n-1) - T_x} \sqrt{\frac{1}{2}n(n-1) - T_y}}$$

Onde:

$T_x = \frac{1}{2} \sum t(t-1)$, sendo t o número de observações empatadas em cada grupo de empates da variável X .

$T_y = \frac{1}{2} \sum t(t-1)$, sendo t o número de observações empatadas em cada grupo de empates da variável Y .

29

2- Se os n indivíduos (elementos) constituem uma amostra aleatória de alguma população, pode-se comprovar se o valor observado τ indica existência de associação entre as variáveis X e Y na população.

H_0 : não existe correlação entre X e Y

H_1 : existe correlação entre X e Y

O método depende do tamanho de n .

Caso1: para $n \leq 10$ a Tabela **Q** dá a probabilidade (unilateral) associada a um valor tão grande quanto um $|S|$ observado.

30

caso 2: para $n > 10$ podemos usar a aproximação normal com

$$Z = \frac{\tau}{\sqrt{\frac{2(n+5)}{9n(n-1)}}}$$

Para os dois casos, obtenha o p-valor $\hat{\alpha}$. Caso $\hat{\alpha} < \alpha$, rejeite H_0 , caso contrário, aceite.

31

Coefficiente de correlação por postos de Spearman

Denominamos $-1 < r_s < 1$ o coeficiente de correlação por postos de Spearman.

É necessário que as variáveis apresentem nível de mensuração, no mínimo, ordinal.

32

Procedimento

Sejam os n pares $(x_1, y_1), \dots, (x_n, y_n)$

1- Atribua postos de 1 a n para cada uma das variáveis X e Y , de modo a formar n pares de postos.

2- Para cada par de postos de X e Y , determine d_i o valor da diferença entre o posto em X_i e o posto em Y_i , $i = 1, \dots, n$

3- O coeficiente de correlação por postos de Spearman é dado por:

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n}$$

33

Exemplo

Calcule o coeficiente de correlação por postos de Spearman para o exemplo anterior (habilidades em matemática e música)

Disciplina	Aluno									
	A	B	C	D	E	F	G	H	I	J
Matemática (X)	7	4	3	10	6	2	9	8	1	5
Música (Y)	5	7	3	10	1	9	6	2	8	4

34

Comentários

1- Em caso de haver observações empatadas, atribuímos a elas a média dos postos que lhe caberiam se não houvesse empate. Nesse caso, o valor de r_s seria modificado da seguinte forma

$$r_s = \frac{\sum x^2 + \sum y^2 - \sum d_i^2}{2\sqrt{\sum x^2 \sum y^2}}$$

Onde:

$$\sum x^2 = \frac{n^3 - n}{12} - \sum T_x \quad \text{e} \quad \sum y^2 = \frac{n^3 - n}{12} - \sum T_y$$

35

$T_x = \sum \frac{t^3 - t}{12}$, sendo t o número de observações empatadas em cada grupo de empates da variável X .

$T_y = \sum \frac{t^3 - t}{12}$, sendo t o número de observações empatadas em cada grupo de empates da variável Y .

36

2- Se os n indivíduos (elementos) constituem uma amostra aleatória de alguma população, pode-se comprovar se o valor observado r_s indica existência de associação entre as variáveis X e Y na população.

H_0 : não existe correlação entre X e Y

H_1 : existe correlação entre X e Y

O método depende do tamanho de n .

Caso1: para $n = 4$ até 30 a Tabela **P** dá os valores críticos de $|r_s|$ para níveis de significância 0,05 e 0,01 (Teste unilateral)

37

caso 2: para $n \geq 10$ pode-se determinar a significância de um valor tão grande quanto o r_s observado, usando a aproximação pela distribuição t de Student. usar a aproximação normal com

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}} \sim t_{n-2}$$

Para os dois casos, obtenha o p-valor $\hat{\alpha}$. Caso $\hat{\alpha} < \alpha$, rejeite H_0 , caso contrário, aceite.

38
